Strategy Selection Versus Strategy Blending: A Predictive Perspective on Single- and Multi-Strategy Accounts in Multiple-Cue Estimation

Herzog, Stefan M ; von Helversen, Bettina

Abstract: The claim that a person can use different strategies or processes to solve the same task is pervasive in decision making, categorization, estimation, reasoning, and other research fields. Yet such multi-strategy approaches differ widely in how they envision that the different strategies are coordinated and therefore do not represent one unitary approach. Toolbox models, for example, assume that people shift from one strategy to another as they adapt to specific task environments based on past experience. Unlike such multi-strategy selection approaches, multi-strategy blending approaches assume that the outputs of different strategies are blended into a joint, hybrid response (i.e., "wisdom of strategies" in one mind). The goal of this article is twofold. First, we discuss strategy blending as a conceptual alternative to strategy selection for modeling human judgment. Second, we investigate the predictive performance of the different approaches in synthetic and real-world environments. Taking a normative perspective, we study the coordination of rule-based and exemplar-based processes in estimation tasks. Our simulations using synthetic and real-world environments indicate that, for medium-sized samples, multi-strategy blending approaches lead to more accurate estimates than relying on a single strategy or selecting a strategy based on past experience—possibly because neither rule- nor exemplar-based processes in isolation are sufficient to capture statistical regularities that enable accurate estimates. This suggests that multi-strategy blending approaches can be advantageous to the degree that they rely on qualitatively different strategies.

DOI: https://doi.org/10.1002/bdm.1958

Reference:

Herzog, S. M., & von Helversen, B. (2018). Strategy Selection Versus Strategy Blending: A Predictive Perspective on Single-and Multi-Strategy Accounts in Multiple-Cue Estimation. *Journal of Behavioral Decision Making*, *31*(2), 233-249. doi: 10.1002/bdm.1958

Strategy Selection versus Strategy Blending:
A Predictive Perspective on Single- and Multi-Strategy Accounts in Multiple-Cue Estimation

Stefan M. Herzog

Max Planck Institute for Human Development, Berlin, Germany

Bettina von Helversen

University of Zurich, Switzerland

Author Note

Stefan M. Herzog, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; Bettina von Helversen, Department of Psychology, University of Zurich, Switzerland.

We thank Susannah Goss for editing the manuscript, and the Swiss National Science Foundation for a grant to the first author (100014_129572/1) and a grant to the second author (100014_146169).

Correspondence concerning this article should be addressed to Stefan M. Herzog, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: herzog@mpib-berlin.mpg.de

Abstract

The claim that a person can use different strategies or processes to solve the same task is pervasive in decision making, categorization, estimation, reasoning, and other research fields. Yet such *multi-strategy* approaches differ widely in how they envision that the different strategies are coordinated and therefore do not represent one unitary approach. *Toolbox* models, for example, assume that people shift from one strategy to another as they adapt to specific task environments based on past experience. Unlike such multi-strategy *selection* approaches, multi-strategy *blending* approaches assume that the outputs of different strategies are blended into a joint, hybrid response (i.e., "wisdom of strategies" in one mind). The goal of this article is twofold. First, we introduce *strategy blending* as a conceptual alternative to strategy selection for modeling human judgment. Second, we investigate the predictive performance of the different approaches in synthetic and real-world environments. Taking a *normative perspective*, we study the coordination of rule-based and exemplar-based processes in estimation tasks. Our simulations using synthetic and real-world environments indicate that, for medium-sized samples, multi-strategy blending approaches lead to more accurate estimates than relying on a single strategy or selecting a strategy based on past experience—possibly because neither rule- nor exemplar-based processes in isolation are sufficient to capture statistical regularities that enable accurate estimates. This suggests that multi-strategy blending approaches can be advantageous to the degree that they rely on qualitatively different strategies. (231 words)

Keywords: strategy selection; strategy blending; hybrid models; judgment; estimation; decision making

Introduction

The claim that a person can use different strategies or processes to solve the same task is pervasive in decision making, categorization, estimation, reasoning, and other research fields (see, e.g., Marewski & Link, 2014; Scheibehenne, Rieskamp, & Wagenmakers, 2013, for reviews and discussion). In decision making, for example, multi-strategy approaches assume that people have a variety of decision strategies at their disposal and that they select among those strategies depending on task affordances (Gigerenzer & Goldstein, 1996; Rieskamp & Otto, 2006). Similarly, estimation and categorization research assumes that people recruit various strategies, including exemplar-based strategies, procedural strategies, and rule-based strategies (e.g., Anderson & Betz, 2001; Erickson & Kruschke, 1998; Juslin, Karlsson, & Olsson, 2008; Nosofsky, Palmeri, & McKinley, 1994).

Multi-strategy approaches differ widely in how they envision that different strategies are coordinated and therefore do not represent one unitary approach. *Toolbox* models (e.g., Gigerenzer & Selten, 2001), for example, assume that people have different strategies at their disposal (i.e., the "toolbox") and that they shift from one strategy to another as they adapt to specific task environments—gravitating towards strategies that are likely to succeed in the task at hand (Gigerenzer & Selten, 2001; Payne, Bettman, & Johnson, 1993; Rieskamp & Otto, 2006; Scheibehenne et al., 2013). Such toolbox models have been proposed in decision making (e.g., Gigerenzer & Selten, 2001; Payne et al., 1993; Rieskamp & Otto, 2006; Scheibehenne et al., 2013) as well as in categorization and estimation research (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Juslin et al., 2008; Nosofsky et al., 1994; von Helversen & Rieskamp, 2008). Some authors assume that a decision maker can, for example, use the take-the-best strategy for one object or task and a weighted additive strategy for the next, with no intermediate state between the two strategies being possible (Bröder & Schiffer, 2003; 2006; Rieskamp & Otto, 2006; Söllner, Bröder, Glöckner, & Betsch, 2014). Similarly, the categorization model COVIS assumes that implicit procedural and explicit rule-based processes "race" for an answer, with the faster one determining the response (Anderson & Betz, 2001; Ashby et al., 1998). In contrast to toolbox models, which insist on shifting from one strategy to the next, *hybrid* models (e.g., Erickson & Kruschke, 1998) allow not only for opportunistic switching between processes, but also for the *blending*[1] of

[1] Note that we use the term blending as short hand for averaging the outputs of more than one strategy into one joint response. This use of the term should thus not be confused with, for example, blended retrieval of chunks or the creation of new production rules (i.e., production compilation) in ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004; Thomson, Lebiere, Anderson, & Staszewski, 2015).

different strategies' outputs into a joint, hybrid response—for example by taking a weighted average of the strategies' outputs as a response (i.e., "wisdom of strategies" in a single mind).

The goals of this article are twofold: First, on a conceptual level, we discuss the advantages and disadvantages of *strategy blending* as an alternative multi-strategy approach for modeling human judgment beyond strategy selection. Second, we investigate the *predictive* performance of various coordination schemes—strategy selection, strategy blending, and others—in estimating real-world quantities (e.g., National Football League scores). More specifically, based on empirical evidence that people use exemplar-based and rule-based strategies when estimating quantities (e.g., Juslin et al., 2008; von Helversen & Rieskamp, 2009), we compare—across a set of synthetic and real-world data sets—the predictive accuracy of relying on one of these strategies, selecting the strategy that performed better in the past, or blending the output of the strategies into a joint response. Building on descriptive research of actual human judgment (see our literature review below), we thus take a *normative perspective* and study, by means of simulation, how successful these coordination schemes are in predicting external, real-world criteria.

Assuming that human cognition is an adaptation that evolved to solve real-world problems (e.g., Anderson, 1990; Chater & Oaksford, 1999; Todd, Gigerenzer, the ABC Research Group, 2012), a normative approach is an important complement to the descriptive study of human behavior. It can inform descriptive approaches by showing how well specific strategies fare in real-world tasks and can help to generate hypotheses—for example, about when humans will rely on which strategies. Here, we follow the approach taken by Gigerenzer and colleagues (Gigerenzer, Todd, ABC Research Group, 1999; Todd et al., 2012) and focus on judgment accuracy as the criterion of interest. In the following, we first describe the estimation strategies and coordination schemes in more detail and then report simulations comparing their predictive accuracy in synthetic and real-world environments.

Models of Estimation and Categorization

Judging quantities, categorizing objects, and making decisions are all crucial tasks for successful human behavior. People executing these tasks use the attributes of an object—such as the features of a digital camera—to infer its value on an unknown criterion (*estimation*: e.g., the camera's worth), classify it to a category (*classification*: e.g., compact vs. bridge camera), or decide among objects (*decision*: choose which camera to buy). A vast and diverse literature in cognitive science and judgment

and decision making has investigated how people achieve these tasks (e.g., Ashby & Maddox, 2005; Gigerenzer, Hertwig, & Pachur, 2011; Kruschke, 2008; Payne et al., 1993).

The different models and strategies proposed can be broadly classified into two classes with respect to the cognitive processes they assume (Hahn & Chater, 1998): First, *similarity-* or *exemplar-based processes* use the similarity to previously encountered cases to make estimates, categorizations, and decisions. Second, *rule-based processes* rely on previously abstracted rules that define the relationship between a specific piece of information and the criterion. There is clear evidence that the same people rely on both kinds of processes and that which of the two processes they use depends on the task. For example, exemplar-based processes are more prevalent than rule-based processes when people do not know which cue values indicate a specific category or a high criterion value (e.g., whether high or low blood pressure should be considered more dangerous; Bröder, Newell, & Platzer, 2010; Newell, Weston, Tunney, & Shanks, 2009; Platzer & Bröder, 2013; von Helversen, Karlsson, Mata, & Wilke, 2013), when only a small number of exemplars is known (Homa, Proulx, & Blair, 2008; Rouder & Ratcliff, 2006), or when the criterion is a nonlinear function of the cues (Hoffmann, von Helversen, & Rieskamp, 2013; 2014; Juslin et al., 2008)—and vice versa. Most research on exemplar- and rule-based processes in estimation and categorization assumes that people switch between strategies (Juslin et al., 2008; Nosofsky et al., 1994; von Helversen & Rieskamp, 2009)—that is, for any object, a response is determined by either a rule or an exemplar-based strategy, but never by both. For example, the Rule-plus-Exception model (RULEX) assumes that people first try simple one-dimensional rules and that they only recruit exemplar-based processes to store exceptions from such rules (Nosofsky et al., 1994).

Hybrid models, in contrast, allow for a combination of both processes. For example, ATRIUM (Erickson & Kruschke, 1998; Vanpaemel & Storms, 2008) assumes that people have two "experts" in their mind: an exemplar-based and a rule-based module, whose outputs are processed by a gating mechanism. This gating mechanism can select between these modules or blend their outputs by averaging their responses. In addition, ATRIUM can learn to rely more strongly on the more successful module in terms of the probability of selecting one module over the other or in terms of the weight put on each model when blending them. This can be done either for the whole task or depending on the object to be categorized (allowing people to learn, for example, exceptions to rules). Thus, ATRIUM not only enables people to rely exclusively on rule-based or exemplar-based processes, but also allows intermediate solutions.

Whether people switch between or blend strategies is still an open empirical question. As reviewed above, extensive research within the strategy-selection framework has provided evidence that the strategy that best describes peoples' estimation or categorizations depends on the task. However, most of this research did not test the strategy-selection account against a strategy-blending account. Nevertheless, there is some evidence indicating that exemplar- and rule-based processes simultaneously influence how humans categorize (e.g., Brooks & Hannah, 2006; Erickson, 2008; Hahn, Prat-Sala, Pothos, & Brumby, 2010) or make estimates (von Helversen, Herzog, & Rieskamp, 2014) based on multiple cues. These findings are consistent with a strategy-blending but not with a strategy-selection account. For example, Hahn and colleagues (2010) found that—even when instructed to use a rule—participants categorized new cases that were similar to previous cases faster than cases that were dissimilar to previous cases, and made fewer errors doing so. Similarly, von Helversen and colleagues (2014) found that the similarity to previously seen candidates influenced how positively job candidates were judged, even though participants combined information about the candidates' résumés in a rule-based manner.

Here, we use a simulation approach to investigate whether selecting or blending rule-based and exemplar-based processes is more advantageous in estimation problems, that is, in predicting real-world criteria from a set of cues. In the following, we first describe how we implemented exemplar-based and rule-based processes on a computational level (Marr, 1982) and then discuss the advantages and disadvantages of strategy selection and strategy blending.

Exemplar model

To represent an exemplar-based estimation process, we used an exemplar model for multiple-cue estimates (Juslin et al., 2008) that extends the generalized context model (GCM; Nosofsky, 1984) to estimation. The model assumes that an estimate $\hat{y}$ of the criterion value of a new object is based on the similarity of the object p to every exemplar i stored in memory, where the estimate $\hat{y}$ is an average of the criterion values x_i of the stored exemplars weighted by their similarity to the target object $S(p, i)$:

$$\hat{y}_p = \frac{\sum_{i=1}^{I} S(p,i) \cdot x_i}{\sum_{i=1}^{I} S(p,i)}, \qquad (1)$$

where *I* is the total number of exemplars stored in memory. The similarity *S(p, i)* between an object and an exemplar is assumed to be a nonlinear function of the distance *d* between the two objects,

$$S(p,i) = e^{-d(p,i)}, \quad (2)$$

where *d* is a function of the difference between the objects' values on each attribute dimension $c_1 \ldots c_j$, the importance of each cue dimension measured by an attention parameter *s*, and a sensitivity parameter *h* that reflects the discriminability in psychological space (Nosofsky & Zaki, 1998):

$$d(p,i) = h\left[\sum_{j=1}^{J} s_j \left| c_{pj} - c_{ij} \right|\right] \quad (3)$$

Two main implementations of the exemplar model are used in the literature: (a) a full version that allows the attention given to each cue to differ and (b) a simplified version with a single free parameter determining the similarity gradient, thus assuming that *s* is the same for all cue dimensions. In support of the full version, it has been argued that exemplar models with unequal attention weights provide a more adequate representation of exemplar-based processes in human categorization, estimation, and decision making (e.g., Anderson & Betz, 2001; Erickson & Kruschke, 1998; Juslin et al., 2008; Rehder & Hoffman, 2005). However, other researchers have argued that a core conceptual difference between rule-based and similarity-based cognitive processes is that rule-based processes focus on a subset of the available dimensions, whereas similarity-based processes spread their attention across more dimensions (Hahn et al., 2010; Milton, Longmore, & Wills, 2008; Pothos, 2005). In line with the latter conjecture, exemplar models with a single attention parameter frequently outperform models with free attention parameters in generalization tests in the context of estimation tasks (e.g., Hoffmann et al., 2013; von Helversen & Rieskamp, 2008; 2009). That is, when a model is fitted to one set of estimates of a participant and the fitted parameters are then used to predict her estimates for a new set of cases, the equal-attention exemplar model describes the new estimates better than the more complex, unequal-attention model does. These results suggest that the attention weights estimated by a full exemplar model will often be too extreme in estimation tasks and that attention is actually spread more evenly across the cues. For this reason and to emphasize the conceptual differences between the rule- and exemplar-based processes, we focus here on the simplified version of the exemplar model. However, in the general discussion we consider to what extent our results may generalize to exemplar models with unequal attention weights.

Rule model

Rule-based strategies represent a broad category of strategies that can differ in the complexity of the assumed rule-based processes. One universal characteristic of rule-based strategies is that they embody a pre-established action routine that can be applied to new cases. This action routine is represented as an abstracted set of rules and therefore no longer requires knowledge of concrete past experiences (i.e., exemplars). Beyond this universal characteristic, there seems to be no clear consensus in cognitive psychology about which strategies can be referred to as "rule-based." In categorization research, rule-based strategies usually refer to *if–then* rules that involve one or two dimensions and are easily verbalized. In estimation research, in contrast, the predominant rule-based strategies are cue abstraction strategies that learn the importance of several cues, rather than just one or two (Juslin et al., 2008; Pachur & Olsson, 2012). Such cue abstraction strategies assume that, for each cue, people abstract a weight reflecting its importance for estimation.

In line with previous research on multiple-cue estimation, we computationally represent a rule-based estimation process using a cue-abstraction model implemented as a multiple linear regression model, which optimizes the cue weights in the learning sample. Such linear additive models have been widely and successfully used to approximate human judgment in a broad range of domains (for a summary, see Brehmer, 1994). Although linear models could, in principle, incorporate interactions, we limited the rule model to consider only main effects, because people do not generally seem to consider interactions when making judgments based on multiple cues in a rule-based manner (e.g., Brehmer, 1994).

In the rule model, the criterion value $\hat{y}$ for a new object p is estimated by summing the weighted cue values of object p:

$$\hat{y}_p = k + \sum_{1=j}^{J} w_j \cdot c_{jp} \qquad (4)$$

where w_j denotes the weight given to the J cue values $c_{1,} \ldots c_j$ of object p; k is the intercept.

Strategy Selection versus Strategy Blending: A Normative Perspective

Drawing on multiple strategies allows a decision maker to adapt to a particular environment by using the strategy that performs best (e.g., Rieskamp & Otto, 2006; Todd et al., 2012). In estimation, if the criterion is a nonlinear function of the cues, people use exemplar-based strategies because they lead

to more accurate judgments than rule-based cue abstraction strategies in these environments. If, in contrast, the criterion is a linear function of the cues, people use cue abstraction strategies, which should be as accurate as—or more accurate than—exemplar-based strategies (Juslin et al., 2008; Pachur & Olsson, 2012; von Helversen & Rieskamp, 2008). Accordingly, if there are both nonlinear and linear environments—and exemplar-based strategies are therefore superior to cue-abstraction strategies in some environments and vice versa in others—choosing the better strategy should lead to higher performance across multiple environments than using the same single strategy in every environment—provided that the ability to identify the better strategy (e.g., based on learning experience) is above chance.

When Is Strategy Selection and When Is Strategy Blending More Accurate?

Whether and when strategy selection or strategy blending is more accurate is a more complex issue. To appreciate the conditions under which blending strategies improves or decreases accuracy relative to selecting a strategy, it is useful to view blending strategies as applying the concept of the "wisdom of crowds" (Page, 2007; Surowiecki, 2004) to a single mind (Herzog & Hertwig, 2014a). Blending estimates or decisions from different people or algorithms, and the conditions under which blending enhances or decreases accuracy, has been discussed in fields ranging from psychology and cognitive science to judgment and decision making, management science, economics, biology, statistics, artificial intelligence, and machine learning (e.g., Armstrong, 2001; Brown, Wyatt, Harris, & Yao, 2005; Davis-Stober, Budescu, Dana, & Broomell, 2014; Grofman, Owen, & Feld, 1983; Hastie & Kameda, 2005; J. Krause, Ruxton, & Krause, 2010; Kuncheva, 2004; Larrick & Soll, 2006; Larrick, Mannes, & Soll, 2012; Lee, Zhang, & Shi, 2011; Luan, Katsikopoulos, & Reimer, 2012; Marling, Sqalli, Rissland, Munoz-Avila, & Aha, 2002; Page, 2007; Timmermann, 2006). More recently, the study of the wisdom of crowds has been extended to the "inner crowd" (Herzog & Hertwig, 2014a), where several judgments from the same person are blended (Herzog & Hertwig, 2009; 2014b; Vul & Pashler, 2008).

Averaging diverse sources (e.g., the forecasts of different experts) leads to higher accuracy than randomly selecting among the sources—as long as the sources make different errors that therefore cancel each other out, at least partly (Larrick & Soll, 2006). If, on the other hand, the errors are similar, averaging will barely increase accuracy over randomly choosing among the sources. The diversity of errors is not the sole driver of the benefits of averaging, however. All other things being held constant, averaging sources outperforms selecting among them if the differences in accuracy among the sources

are small and it is difficult to reliably select the better source in advance (see Herzog & Hertwig, 2014b; Soll:2009ie, see also Mannes, Soll, & Larrick, 2014). Applying these insights to the multi-strategy framework helps to delineate the conditions under which blending strategies will improve or decrease accuracy relative to selecting a strategy: It will depend on (a) the difference in accuracy between the two strategies, (b) the ease or difficulty in reliably choosing the better strategy in advance, and (c) the degree to which the individual strategies make similar or different errors. We discuss these three points in turn.

First, the difference in accuracy between exemplar- and rule-based strategies will depend on how well the statistical structure of the environment can be represented by either strategy in isolation. The linear rule model we are using can well approximate a linear additive function of cues, but it cannot approximate a multiplicative function. In contrast, an exemplar model can well approximate a multiplicative function of cues. Furthermore, although an exemplar model can also approximate a linear additive function of cues, a linear rule model does so better. Fundamentally, a strategy-selection approach presupposes that one of the strategies in the toolbox approximates the environment so well that it makes sense to prefer it to the others on account of its superior accuracy. Given that rule-based and exemplar-based strategies differ in how well they approximate linear and multiplicative structures, this raises two questions: (1) Do estimation environments differ strongly enough with respect to those statistical characteristics (or do they rather consist of a mixture of linear and multiplicative components)? (2) Is the accuracy gain of trying to approximate a particular statistical structure large enough to justify changing the estimation strategy? To the extent that the statistical structure of the environment cannot be approximated by either an exemplar- or a rule-based strategy in isolation, but possibly by a blend of strategies (see Dietterich, 2000), the difference in accuracy between the strategies might not be consequential enough to outperform blending.

Second, the ease or difficulty in reliably choosing the better strategy in advance will depend on the amount of experience with the environment and the degree to which one strategy is superior to the others (see above). It will often be unclear in advance which strategy is superior; this knowledge therefore has to be acquired by experience. Selecting a strategy based on a—possibly small—learning sample of past experiences bears the risk of misidentifying the structure of the environment (linear vs. multiplicative) and thus betting on the wrong strategy (rule- vs. exemplar-based). The question is therefore whether a sufficiently large learning sample has been obtained to allow the sufficiently reliable selection of a strategy. Blending strategies may often be "biased" in the sense that it may not be able to learn the true underlying function—if the environment happens to be purely linear or purely

multiplicative—but blending will also often be more reliable, and thus more robust, than strategy selection because it hedges against the risk of selecting the wrong strategy (see also the "bias–variance dilemma;" Brighton & Gigerenzer, 2015; Geman, Bienenstock, & Doursat, 1992; Geurts, 2010; Gigerenzer & Brighton, 2009).

Finally, to understand the degree to which the individual strategies will make similar or different errors, it is instructive to consider the statistical truism that averaging can improve accuracy by canceling out opposing biases, as well as nonredundant unsystematic errors (e.g., Larrick & Soll, 2006; Soll, 1999). To the extent that neither strategy can well approximate the statistical structure of an environment (see above), there is always the possibility that the strategies exhibit different biases because they rely on different representations (exemplars vs. rules), which may approximate different aspects of the environment. If so, the strategies' biases will, at least partly, cancel each other out when blending. Irrespective of the fit to the environment, learning a strategy (e.g., the cue weights in the rule model) based on possibly limited experience is already a formidable challenge in itself. Because the reliability of strategies increases (i.e., the size of unsystematic errors decreases) when they are trained on increasingly larger learning samples, strategies should profit from blending, especially when learning experience is limited.

The Complementary Nature of "Lazy" and "Eager" Learning Approaches

One reason why exemplar and rule models might benefit from a blending approach is that they embody fundamentally different approaches to estimation, classification, and decision making. Whereas exemplar models belong to the class of "lazy" learning algorithms, rule models belong to the class of "eager" learning algorithms. "Lazy" models such as exemplar models do not induce any knowledge during learning; they merely store the instances they encounter. It is only when they are prompted with a question that they compute their answers based on relevant instances retrieved from memory. This kind of approach is very flexible because the task to be solved does not need to be known during learning (e.g., Aha, 1997; Juslin & Persson, 2002). "Eager" learning algorithms such as rule models, in contrast, abstract knowledge during learning. The task therefore needs to be specified during learning; otherwise it remains unclear what knowledge should be abstracted. For example, a cue abstraction model needs to know which variable is the criterion variable and which are the cues; it is otherwise unclear which cue weights should be learned.

A strictly "lazy" approach also has its drawbacks, however (Aha, 1997), especially when considering real-world environments. First, "lazy" models need to retain the learning data as their knowledge base so that they can compute lazy answers later on. In contrast, "eager" approaches do not need to retain the learning data; they have already abstracted the relevant knowledge. Second, the flexibility of "lazy" models (in terms of the functional forms they can learn) can make them suffer from more variance (Geman et al., 1992; Geurts, 2010) than "eager" models (e.g., Briscoe & Feldman, 2011). That is, a "lazy" model's inference for a particular new object can change dramatically due to small changes in the learning data (i.e., overfitting). "Eager" models, in contrast, are typically less flexible, that is more biased, than "lazy" models because they necessarily have a priori constraints on what they can learn. This is because learning rules or, more generally, abstracting knowledge presupposes a representational structure according to which experience is turned into abstracted knowledge (e.g., linear relations between cues and criterion in linear models). As the bias–variance dilemma shows, being constrained or biased can be a good thing for predictive performance (Brighton & Gigerenzer, 2015; Gigerenzer & Brighton, 2009).

In sum, "lazy" and "eager" learning algorithms both have their advantages and disadvantages. We argue that these advantages and disadvantages typically complement each other and that a blending approach is one way to harness this complementarity.

Overview of Simulation Studies

We conducted two simulation studies to compare the benefits of strategy selection and strategy blending in estimation tasks. In a first step, we constructed three synthetic environments, in which we manipulated whether the criterion was a linear or a multiplicative function of the cues or a mixture of both. The exemplar-based strategy should be more accurate than the rule-based strategy in the multiplicative environment, whereas the rule-based strategy should outperform the exemplar-based strategy in the linear environment (e.g., Juslin et al., 2008). The simulation should thus provide some insight into how much benefit can be expected from (a) drawing on two strategies rather than relying on a single strategy and (b) using strategy blending instead of strategy selection. In a second step, we used a collection of five real-world data sets to investigate whether the results can be generalized to real-world environments and whether characteristics of those environments influence which multi-strategy approach is more accurate.

Simulation Study I:
Performance of Strategy Selection versus Strategy Blending in Synthetic Data Sets

Synthetic Data Sets

We created three synthetic data sets: a linear, a multiplicative, and a mixed data set. In the *linear data set*, the criterion is a linear function of the cues; in the *multiplicative data set*, the criterion is a multiplicative function of the cues. We created the *mixed data set* by taking half of the objects from the linear data set and the other half from the multiplicative data set, thus creating a data set that incorporates both linear and multiplicative components.

Each of the data sets consisted of 1,000 objects characterized by six cues and a continuous criterion variable. A summary description of the data sets can be found in Table 1. To create the data sets, we drew 1,000 objects described by six cues from a multivariate normal distribution with a mean of zero; the covariances were drawn from a normal distribution with a mean of zero and a standard deviation of 0.15. We then used the cumulative distribution function of the standard normal distribution to convert the cue values to uniformly distributed values between 0 and 1 and multiplied them by 10 to achieve uniformly distributed cues with values between 0 and 10. This procedure yielded cue values comparable to those used in experimental studies (e.g., Karlsson, Juslin, & Olsson, 2007). For the linear data set, the criterion was created by a linear additive function of the cues (see Hoffmann et al., 2013; 2014; Juslin et al., 2008):

$$Y_l = 3.5c_1 + 3c_2 + 2.5c_3 + 2c_4 + 1.5c_5 + 1c_6, \quad (5)$$

where Y_l denotes the criterion value and c_n the cue values, resulting in deterministic criterion values ranging from 0 to 135. For the multiplicative data set, we transformed the linear criterion to obtain a criterion that was a multiplicative function of the cues, but that had a similar range of criterion values, using the following equation (see also Juslin et al., 2008, for a similar approach):

$$Y_m = \exp(Y_l/15) \text{ x } (\max(Y_l)/\max(\exp(Y_l/15))) \quad (6)$$

Finally, to introduce irreducible uncertainty (Hammond, 1996), we added to each criterion value a random error drawn from a normal distribution with a mean of zero and standard deviation of 20. For the mixed data set, we created the criterion by randomly drawing half of the criterion values from the linear criterion and half from the multiplicative criterion.

Simulation Setup

For each simulation run, we randomly drew a learning sample and a test sample. We then fitted the free parameters of the exemplar and the rule model to the learning sample—minimizing the root mean square error (RMSE) between model predictions and criterion values—and used the estimated parameter values to make predictions for the objects in the test sample (using the six coordination schemes described below). We measured estimation accuracy in the test sample using the RMSE between the model's predictions and the criterion values, a commonly used measure of absolute goodness of fit. In addition, we used seven different sizes of learning samples (20, 40, 60, 80, 100, 200, and 500 objects) to vary the amount of experience within an environment. Previous research has shown that the size of the learning sample can strongly impact how well strategies generalize (Gigerenzer & Brighton, 2009; Perlich, Provost, & Simonoff, 2003). All test samples consisted of 250 objects not included in the learning sample. For each data set and each of size of learning sample, we ran the simulation 1,000 times and summarized the results by averaging.

To facilitate interpretation of the results and comparison across data sets, we standardized performance using the standard deviation (SD) of the criterion variable in a data set (i.e., Root Mean Squared Scaled Error, RMSSE; see also Hyndman & Koehler, 2006).[2] That is, separately for each environment, we divided the RMSE values of every strategy by the SD of the criterion values of all objects in that data set. A value of 1 indicates that a strategy's error equals this SD. A value *below* 1 indicates that a strategy's error is *lower* than the SD. For example, a value of 0.8 means that a strategy's RMSE corresponds to 80% of the SD or, put differently, is 20% smaller than the SD. Correspondingly, a value *above* 1 indicates that a strategy's error is *smaller* than the SD. For example, a value of 1.1 means that a strategy's RMSE corresponds to 110% of the SD or, put differently, is 10% larger than the SD. If, say, strategy A has an RMSSE of 0.80 and strategy B has an RMSSE of 0.81, this means that strategy A outperforms strategy B by one percentage point of the SD of the criterion variable.

Coordinating the Rule- and Exemplar-Based Strategies

We tested six schemes for coordinating rule- and exemplar-based strategies to make predictions for the test sample. The first two schemes used just one of the two strategies exclusively: *rule model* or *exemplar model* (see introduction for a formal description of the two strategies).

[2] We used the standard deviation to scale the errors because it measures the deviation from the grand mean on the same scale as the RMSE.

In addition, we tested two versions of strategy selection and strategy blending by varying the level at which coordination—either selection or blending—takes place: at the task or at the object level. In the ecological rationality and adaptive toolbox approach (Todd et al., 2012), it is implicitly assumed that the selection of strategies happens at the task level. That is, once learning is completed, all decisions within the same task are solved using the same strategy. For instance, strategy selection learning theory (SSL) assumes that people learn which strategy is most successful in a specific task—using reinforcement learning—and then exploit that strategy (Rieskamp & Otto, 2006). In real-world environments, it may often be difficult to decide whether or not a particular object belongs to a known task, but people may learn to use strategies for specific task domains. For example, a doctor may learn to rely on one strategy when diagnosing patients but on another when making financial decisions. However, strategy selection and blending can also occur at the object level (Erickson, 2008; see Erickson & Kruschke, 1998; Jacobs, 1999; Jacobs, Jordan, Nowlan, & Hinton, 1991). Some objects may be better captured by a rule, whereas others require memorization (Nosofsky et al., 1994), leading people to use a rule for one object and an exemplar-based process for another. Similarly, some objects may be better captured by a particular process, whereas others may benefit from a blending of both processes, suggesting that the extent of blending may differ from object to object. To account for coordination at this level, we therefore compared selection and blending at the task and at the object level.

Accordingly, the third and fourth schemes selected either the exemplar or the rule model depending on its anticipated success in the test set (*selection-task* and *selection-object*, respectively). *Selection-task* picked, in each simulation run, the process that was superior in the learning sample and used it for all objects in the test sample. To account for differences in model complexity, we used the Bayesian Information Criterion (BIC; Schwarz, 1978) as a model selection criterion. The BIC is commonly used to compare model fit; it takes model complexity into account by penalizing for the number of free parameters. We calculated the BIC using the approximation by Raftery (1995), which is based on the amount of variance explained by the model (p. 135):

$$\mathrm{BIC}_i = n \times \log(1 - R_i^2) + k_i \times \log(n), \qquad (7),$$

where n denotes the number of observations, R^2 the amount of variance explained, and k the number of free parameters of model i. Thus, the smaller the BIC the more parsimoniously the model captures a participant's estimates.

In contrast, *selection-object* picked, in each simulation run *and for each object in the test sample*, the process that was more likely to be superior for that test object—based on its past performance on similar objects in the learning sample. Specifically, for each test object, we calculated the absolute error *e* that the exemplar and the rule model made on each object *i* of the learning sample *I* and then calculated the expected error $\hat{E}$ as the average error across the learning objects weighted by their similarity *S* to the test object *p* in question:[3]

$$\hat{E}_p = \frac{\sum_{i=1}^{I} S(p,i) \cdot |e_i|}{\sum_{i=1}^{I} S(p,i)} \qquad (8)$$

The similarity between learning and test objects was calculated in the same way as for the exemplar model (see equations 2 and 3), using the parameter values estimated for the exemplar model. The response was determined by the relative expected error of the two models, using Luce's (1959) choice rule as a decision rule:

$$\Pr(p,rule) = 1 - \frac{E(p,rule)}{\sum_{m=1}^{M} E(p,m)}, \qquad (9)$$

with Pr*(p, rule)* indicating the probability that the rule model was chosen for test object *p*, *E* the expected error of the rule model, and *M* the number of models *m* from which a response could be selected (in this case, the exemplar and the rule model).

The fifth and sixth scheme blended the outputs of the exemplar and the rule process to make a joint prediction (*blending-task* and *blending-object*, respectively). At the task level, *blending-task* computed for each test object the arithmetic mean of the predictions of the rule and the exemplar model; that is, it treated all objects the same. At the object level, *blending-object* used in each simulation run *and for each object in the test sample* a weighted average of both models' predictions:

$$\hat{y}_p = \left(1 - \frac{E(p,ex)}{\sum_{m=1}^{M} E(p,m)}\right) \cdot y_{ex} + \frac{E(p,ex)}{\sum_{m=1}^{M} E(p,m)} \cdot y_{rule}, \qquad (10)$$

[3] Note that for a single object, absolute error and RMSE are identical.

where y_p denotes the response for object p, blending the response of the exemplar model (y_{ex}) with the response of the rule model (y_{rule}), and E is the expected error of the models m. The weight each model received for a specific test object was—as in *selection-object*—a function of the past performance of the two models in the learning sample, weighted by their similarity to the test object under consideration (see equation 8).

Results & Discussion

Figure 1 and Table 2 show for the three synthetic data sets the average, standardized generalization performance of the six coordination schemes as a function of the size of the learning sample. Because all schemes performed poorly with the smallest learning sample size (20), and to facilitate comparison across strategies, Figure 1 shows only the results for sample sizes from 40 to 500. The poor performance with small sample sizes is probably due to overfitting.

As expected by the design of the data sets, in the linear data set, the rule model outperformed the exemplar model at every size of the learning sample. The reverse held for the multiplicative data set (compare the left with the right panel of Figure 1), although the advantage of the exemplar model decreased with sample size. The advantage of one model over the other ranged from 1 to 6 percentage points in the linear data set and from 1 to 15 percentage points in the multiplicative data set (see Table 2). In the mixed data set, the predictive accuracy of the model depended on the size of the learning sample (see middle panel of Figure 1). With small samples (40 and 60), the exemplar model was more accurate; with larger samples, the rule model was somewhat more accurate. Taken together, the exemplar model seems more robust but less flexible than the rule model—putting it at an advantage in terms of predictive accuracy in small sample sizes but at a disadvantage in large sample sizes, which allow for reliable estimates of the importance of each cue dimension in the rule model.

Considering all three data sets together, using either strategy selection or strategy blending was superior to consistently relying on just one of the individual strategies (rule or exemplar model). For the most part, the performance of the two selection schemes (task or object level) fell between that of the exemplar model and the rule model. Strategy blending performed even better (and there were almost no differences between blending on the task and object level). In the linear data set, strategy blending performed somewhat worse than the rule model, but clearly better than the exemplar model. In the multiplicative data set, its performance fell between that of the exemplar and the rule model for the two smallest sample sizes and the largest sample size, and it was as accurate as the exemplar model for the

intermediate sample sizes. In the mixed data set, the performance of strategy blending fell between that of the exemplar model and the rule model for the smallest sample size, it outperformed both for sample sizes between 40 and 100, and it performed as well as or only slightly worse than the rule model (the better of the two single strategies) for the two largest sample sizes (200 and 500).

Taken together, these results suggest that blending was the most robust coordination scheme overall: Considering all three environments simultaneously, it performed best. Furthermore, assuming that purely linear or multiplicative environments are rare in the real world, these results suggest that strategy blending should perform well in real-world tasks. To investigate this hypothesis, we conducted a second simulation study using five real-world environments.

Simulation Study II: Performance of Strategy Selection versus Strategy Blending in Real-World Environments

Data Sets

We analyzed a collection of five real-world data sets that has previously been used to compare the performance of proper and improper linear models (Dana & Dawes, 2004; see this reference for original references and data sources) using the same simulation setup as in Study I. We chose this convenience collection of five data sets for three main reasons. First, the data sets stem from different content domains (i.e., biology, sports, public opinion, political sentiment, and occupational prestige). Second, the data sets differ in their statistical structure (e.g., linear predictability and distribution of criterion values), which may affect the relative accuracy of rule- and exemplar-based processes. Third, using a pre-compiled collection of data sets, instead of taking individual data sets from various sources, leaves researchers less room for "cherry-picking."

In all data sets, a continuous criterion variable was predicted by several cues (see below for details). A data set's statistical structure likely influences the strategies' performance profiles. As possible proxies for the functional relationship between criterion and cues, we considered the proportion of linear variance that could be explained by the cues (R^2) and the skewness of the distribution of criterion values. Table 3 presents details of the data sets' statistical structure. Multiplicative relationships between cues and criterion often lead to highly skewed criteria distributions (often called "J-shaped") and linear rules do not work well in such environments (Hertwig, Hoffrage, & Sparr, 2012; von Helversen & Rieskamp, 2008).

In two data sets, the task was to predict peoples' responses based on their characteristics (ABC and NES). In two data sets, the criterion was a measure of success (NFL and WLS); in the final data set, the goal was to estimate a biological magnitude. In the following, we describe the data sets in detail.

- The *Abalone data set* contains 4,177 cases. The criterion—the age of an abalone (a sea snail)—is predicted by seven measurements: shell weight, diameter, height, length, whole weight, viscera weight, and shucked weight.
- The *ABC data set* contains 955 cases from a random polling of U.S. households by ABC News in 2002. The criterion is respondents' confidence that Osama Bin Laden would be captured or killed. The predictors are age, gender, level of education, and whether participants regularly displayed the American flag, and how proud they were to be American.
- The *NES data set* contains 1,910 cases from a telephone poll during the 1988 U.S. presidential primary elections. The criterion is how positively the Republican Party was rated on a scale from 0 to 100. The six predictors are the answers to poll questions asking participants (1) if they thought the nation's economy was better or worse than the year before, (2) if they were financially better or worse off than the year before, and to indicate their agreement with the following statements: (3) "If people were treated more equally in this country, we would have many fewer problems," (4) "Changes in lifestyle, like men and women living together without being married, are signs of increasing moral decay," (5) "We have gone too far in pushing equal rights in this country," and (6) "We should be more tolerant of people who choose to live according to their own moral standards, even if they are very different from our own."
- The *NFL data set* contains 3,057 cases: the outcomes of the National Football League games from 1981 to 1995, excluding strike years. The criterion is the difference in final scores (home team minus visiting team) predicted by 10 team statistics: points per game, points allowed per game, passing rating, interceptions thrown, total yards of offense, total yards allowed, percentage of opponents' plays ending in a sack, opponents' average punt return, opponents' average kickoff return, and percentage of plays penalized.
- The *WLS data set* includes 6,385 cases taken from the Wisconsin Longitudinal Survey (1993). The criterion is occupational prestige in 1992, predicted by (1) a measure of

physical health, (2) number of children and measures of (3) depression, (4) extraversion, and (5) neuroticism.

Results & Discussion

As shown in Figure 2, in most data sets, most models had errors smaller than the standard deviation of the criterion variable (i.e., had RMSSEs < 1) once the size of the learning sample was 40 or larger (see also Table 4). At the largest sample size (500), the decreases in error ranged between 2 and 30 percentage points. The largest decrease was in the Abalone data set; the smallest in the WLS data set. The poor performance with small sample sizes was probably due to overfitting, in particular for the rule model, which had the largest number of free parameters. Furthermore, in some data sets, the criterion could not be predicted very well by the cues (e.g., WLS; see R^2 values in Table 3), making it difficult for the models to have RMSSEs < 1 (an RMSSE > 1 means that using the grand mean of all criterion values as the same estimate for all objects was more accurate than the model's estimates).

Across data sets, the exemplar model outperformed the rule model at small sample sizes (20 and 40), but the rule model caught up and performed as well as or better than the exemplar model at larger sample sizes (Figure 2, summary panel, upper left); this pattern of performance is similar to that observed in the mixed data set in our simulation study (Figure 1, middle panel). The data sets differed with respect to which of the two models performed best. The rule model outperformed the exemplar model in the Abalone and the ABC data sets, whereas the exemplar model outperformed the rule model in the WLS data set. In the NFL and the NES data sets, the exemplar model was more accurate with smaller samples, but the rule model was more accurate with larger samples. This variation does not seem to be related to whether the task was to predict a human response (e.g., ABC: a response in a poll) or a nonsocial criterion (e.g., Abalone: the age of an abalone snail), suggesting that the statistical structure of the task is more important than the domain. Consistent with this conjecture, the average Spearman correlation across data sets between (a) the difference in accuracy between the rule and the exemplar model and (b) the amount of linear variance explained by the cues (R^2)—calculated per sample size and then averaged—was $r_S = .73$, $SD = .31$, which suggests that the rule model had an advantage over the exemplar model in tasks that were linearly predictable.

Comparable to our findings for the synthetic data sets, the performance of the two strategy-selection schemes mostly fell between that of the exemplar and the rule model (i.e., the strategy-selection schemes were more accurate than the inferior of the two individual strategies in any task, but

less accurate than the superior of the two). However, the performance of the two strategy-selection schemes differed among data sets and was sometimes worse than, or as bad as, the inferior of the two individual strategies. This suggests that it was not possible to reliably identify the superior strategy based on learning experience. Nevertheless, it was advantageous to rely on a multi-strategy approach because neither the exemplar model nor the rule model was consistently the best strategy, and even unreliably selecting among strategies was better than not switching at all.

However, the two strategy-selection schemes were outperformed by the two strategy-blending schemes. A simple blend of the two individual strategies (*blending-task*) outperformed both strategy-selection schemes (*selection-task* and *selection-object*) in four of the five data sets (for all sample sizes > 20). Overall, the advantage of strategy blending over strategy selection was relatively stable across sample sizes, whereas the absolute difference in accuracy between the exemplar and the rule model was much larger for small than for large samples. This shows that the advantage of strategy blending over strategy selection was not just a function of the difference in accuracy of the exemplar and rule model; otherwise, the advantage of strategy blending over strategy selection would have shown a similar decline as sample sizes increased. It seems possible that the effect of the large differences in accuracy between the exemplar and rule model for smaller samples was offset by the difficulty in identifying the better strategy with such small samples.

Although strategy blending performed worse than the exemplar model at the smallest sample size (20), when summarized across environments, *blending-task* was about 1 to 2 percentage points better than the superior of the two individual strategies in the range of 40 to 200 training objects and equally good with 500 training objects (see Figure 2, summary panel; and Table 4). Blending outperformed the rule model at medium to small sample sizes in all data sets (i.e., sample sizes < 200); even with large samples, its performance was as good or only slightly worse. In comparison with the exemplar model, blending was less accurate for small samples, but more accurate at sample sizes of 40 or larger in all but one data set. Overall, the differences between data sets were relatively small, with the largest benefit of blending in the Abalone data set (around 4 percentage points for medium-sized samples) and the smallest benefit in the WLS data set, where blending performed only at the level of the exemplar model (which was more accurate than the rule model in this data set).

The advantage of strategy blending was not systematically related to any of the statistical characteristics of the data sets. Spearman correlations across the five real-world data sets and the three synthetic data sets between the performance gain of blending rather than using the better of the two

individual strategies (rule or exemplar model) and either (a) linear predictability or (b) skewness did not show a coherent pattern. Correlations varied strongly with sample size and were, on average, of medium size (linear predictability: $Md = -.42$, interquantile range or $IQR = .52$; skewness: $Md = -.38$, $IQR = .46$). Given the small number of data sets on which this analysis relies, we conclude that there is, at least, no indication of a strong relationship between the statistical characteristics of the data sets and the benefits of strategy blending. Finally, although the advantage of strategy blending was largest in the Abalone data set, it does not seem that the benefits of blending are restricted to nonsocial estimation problems.

Across data sets, there was no clear difference between the task- and object-based coordination schemes. For strategy blending, a simple blend of the strategies (i.e., the simple average of the two strategies' responses; *blending-task*) and a weighted blend of the strategies (i.e., weighted average according to the two strategies' past performance on similar objects; *blending-object*) performed roughly the same. For strategy selection, whether using the same strategy for all objects (*selection-task*) or selecting the strategy according to its past performance on similar objects (*selection-object*) showed the better performance differed among environments.

Which strategies performed best overall? When summarized across the seven sizes of learning samples, the results will depend partly on the particular choices of sample sizes used (20, 40, 60, 80, 100, 200, and 500). Nevertheless, giving an overall impression may be desirable (see Table 4): Blending-task and blending-object were most often among the best-performing strategies (83% and 60% of data-set–sample-size combinations, respectively), whereas all other strategies lagged behind on this measure of success: exemplar model (29%), selection-task (17%), rule-model (14%) and selection-object (0%).

General Discussion

Many cognitive models of estimation, categorization, and decision-making assume that the same person can use both exemplar- and rule-based strategies to solve the same task (e.g., Erickson & Kruschke, 1998). Yet it has remained unclear whether using both strategies provides an advantage over using just one strategy and, when both strategies are available, whether it is better to select a single strategy or to blend multiple strategies—depending on the task or even on the object within the task. Our simulations using synthetic and real-world environments indicate that multi-strategy approaches, such as strategy selection and strategy blending, lead to more accurate estimates than relying on just one strategy

across all environments. One reason for the advantage of multi-strategy approaches seems to be that natural environments are sufficiently heterogeneous with respect to whether rule-based or exemplar-based strategies render more accurate estimates; therefore, using both strategies improves performance.

Should a judge select one strategy or blend the responses of both strategies? Our findings suggest that—with the exception of very small learning samples—selecting the strategy that performed best in the learning sample is *less* accurate than blending the responses of the two strategies. With medium-sized learning samples, strategy blending not only outperformed strategy selection, but was also more accurate than either the exemplar or the rule model by about 2 percentage points. Although, at first glance, this difference may not seem consequential, it is comparable to the differences in accuracy between the rule and the exemplar model across environments and sample sizes. Furthermore, even small increases in performance can be relevant when benefits accumulate over time (e.g., Haldane, 1927).

Why does blending strategies perform so well? The inferior performance of the strategy-selection schemes suggests that it may often be difficult to identify—reliably enough—which of the two strategies would generalize better (see Soll & Larrick, 2009), based on their performance in the learning samples. Blending seems to offer a robust compromise that allows people to benefit from the two estimation strategies' abilities to exploit different aspects of the environment. This, in turn, allows the two estimation strategies to make different kinds of errors, which are—at least partly—cancelled out when their estimates are blended. However, our results also indicate boundary conditions for blending: With very small learning samples, the exemplar model outperformed the rule model to such a degree that incorporating the rule model's estimates by blending decreased performance relative to relying solely on the exemplar model. Similarly, with large samples, blending the exemplar and the rule model did not help much, presumably because both strategies made very similar predictions, leading to highly correlated errors.

Can our findings on the benefits of strategy blending be generalized to other estimation strategies, to categorization strategies, and to other domains of human behavior? The benefits of the multi-strategy blending approach seem to stem from the mixture of statistical structures found in real-world environments, which cannot be captured by a single process in isolation (see Dietterich, 2000), from the nonredundant errors of the two qualitatively different cognitive processes (Herzog & Hertwig, 2009; 2014a), and from the difficulty people evidently have in selecting the—sometimes only slightly—more accurate process based on past experience (Mannes et al., 2014; Soll & Larrick, 2009). This

suggests that our results should generalize to other domains—such as categorization, decision making, or problem solving (e.g., Page, 2007)—to the degree that these conditions are fulfilled.

Cognitive Modeling and Predictive Simulations

The goal of our simulations—using stylized estimation strategies—was to provide an existence proof that people might benefit not only from adaptively switching between estimation strategies, but also from blending them. Such a strategy blending approach can provide accuracy benefits that go beyond exploiting the fact that different strategies work best in different environments (Todd et al., 2012). One question, however, is to what extent our simulation results, which relied on one specific implementation each of a rule- and an exemplar-based process, can be generalized to human estimation processes more generally. The answer depends on whether our model implementations captured the relevant aspects of human estimation and learning.

One question is whether implementing the exemplar model with *unequal* attention weights would have been more appropriate than the equal-attention version we used. As discussed above, if human participants learned to solve the tasks we simulated using an exemplar approach, their attention weights would probably fall in between the equal weights we assumed and the more dispersed, optimized attention weights that a model with free parameters would estimate. This, in turn, suggests that fitting an exemplar model with free attention parameters would overestimate the differences in attention across cues, possibly more so than using equal weights would underestimate the respective differences. One solution for future research would be, instead of using optimized attention weights, to specify learning models that can capture how people learn to allocate attention. Although some learning models exist in categorization (Erickson & Kruschke, 1998; Kruschke, 1992), there is as yet relatively little such work in estimation (for notable exceptions, see Kelley & Busemeyer, 2008; Speekenbrink & Shanks, 2010). Such specifications would allow for more accurate simulation of the learning and selection processes in multi-strategy accounts of human judgment.

Nevertheless, it seems worthwhile to discuss how our results would be affected by using a more flexible exemplar model. For one, the higher generalization performance of the exemplar model—relative to the rule model—when trained on only small learning samples is probably partly due to the constraints imposed by its simplicity. That is, an exemplar model with a separate, free attention parameter for each cue dimension would probably overfit with small samples, but might generalize better with larger samples. In addition, there is the possibility that using an exemplar model with free

attention parameters would decrease the advantage of strategy blending over strategy selection. To the extent that both models are able to estimate the importance of a specific cue dimension accurately from the learning sample, this could increase the correlation between predictions and thus make blending less successful. However, because exemplar-based processes differ from rule-based processes in representation and information processing (Hahn & Chater, 1998; Johansen & Palmeri, 2002), richer, more cognitive implementations of such models should arguably also reflect said differences—possibly even more so. Accordingly, using such more cognitive implementations may also accentuate error cancellation through blending and thus amplify the results found in the present study.

Environmental Structure and Multi-Strategy Accounts of Judgment

In the mixed environment consisting of a mixture of linear and multiplicative components, we found that the exemplar model was more accurate for small sample sizes, but that the rule model was more accurate for larger sample sizes. In the real-world data sets, the degree to which the rule model outperformed the exemplar model correlated positively with the proportion of linear variance explained by the cues in a data set. Although a low linear predictability could also indicate that the cues simply did not allow for accurate prediction of the criterion—either linearly or nonlinearly—the result is consistent with the conjecture that linear predictability is a key characteristic distinguishing between environments in which the rule model outperforms the exemplar model and vice versa.

Real-world environments might not consist of purely linear or multiplicative components, however, but might represent a mixture of both components. The difference in performance between the exemplar and rule model across the real-world data sets was smaller than that in the synthetic linear and multiplicative data sets, which is consistent with what one would expect if environments are not purely linear or multiplicative. Furthermore, overall, the pattern of performance in the real-world data sets was most similar to that in the mixed synthetic data set, which is also consistent with the conjecture that real-world environments are often not purely linear or multiplicative. If this were indeed the case, it would render strategy blending successful in real-world environments because neither strategy in isolation can capture the statistical structure of the environment (Dietterich, 2000). Consistent with this conjecture, blending was more successful in the mixed synthetic data set and in the real-world data sets than in the purely linear and purely multiplicative synthetic data sets. However, we must emphasize that this is currently only a conjecture. Our simulations can only provide tentative answers given (a) that it is not clear how to best measure linear and multiplicative components in natural data sets and (b) that we

investigated only a handful of environments; a collection of a few dozen environments would seem necessary to derive reliable conclusions.

Although the real-world data sets varied in many respects—including their linear predictability, the relative accuracy of the rule and exemplar model, the number of cues, the skewness of the criterion variable, and whether the criterion was human response or natural observations—we found relatively little variance in the benefits of strategy blending across data sets. The Abalone data set was most distinct in that it was best predicted by a linear model, but was also most skewed and was the only completely nonsocial data set. However, the benefits of strategy blending were not restricted to nonsocial data, and our simulation with synthetic data sets suggests that statistical characteristics may be more important than the domain content in determining which strategy works best. Further research on the statistical features that render strategy selection or strategy blending more successful is necessary.

Psychological Insights and the Merits of a Normative Perspective

At first sight, blending the responses of different processes or strategies may not seem a parsimonious approach because it requires that at least two processes are adopted simultaneously. However, this assumption is shared by many strategy-selection accounts. Whereas in strategy selection the better or faster response is selected and the other response is ignored, in strategy blending both responses are integrated. Furthermore, the idea that people rely on both exemplar- and rule-based processes is in line with both the empirical evidence and, more broadly, the historical development in the literature on computational modeling approaches, from models positing a single process (e.g., Kruschke, 1992; Nosofsky, 1984; Nosofsky & Johansen, 2000) to models integrating rule- and exemplar-based processes. In addition, empirical evidence suggests that exemplar similarity influences responses even if a rule suffices to solve the task and if people clearly rely on a rule when making their judgments (Brooks & Hannah, 2006; Hahn et al., 2010; von Helversen et al., 2014)—a result that is implied by a strategy blending approach, but is difficult to reconcile with strategy selection accounts. These descriptive results resonate with our results, which suggest that in many real-world environments it is beneficial to rely on both processes at the same time. In addition, our results suggest that strategy blending could be particularly useful with medium-sized learning samples. One reason could be that, in these situations, people have experienced some exemplars that they can retrieve from memory, but have not yet gained enough experience to abstract a reliable, successful rule. However, more empirical and modeling research is necessary to understand when and how people might blend the responses from

exemplar- and rule-based processes. Furthermore, we implemented blending as a simple average of the two strategies' responses, which ignores learning processes. Future research should consider psychologically plausible implementations of how the responses of two strategies are blended. For example, it would be interesting to implement the blending of exemplar- and rule-based processes within a unitary framework such as ACT-R (Anderson, 1990; Anderson et al., 2004; see also Anderson & Betz, 2001, for an implementation that essentially assumes a race between exemplar and rule-based production rules), using, for example, blended retrieval of chunks or production compilation (i.e., creation of new production rules; Anderson et al., 2004; Thomson et al., 2015).

In general, a normative perspective has merits for research in psychology and judgment and decision making (JDM). Examining the ability of cognitive models to predict real-world criteria goes a step further than comparing their ability to describe human behavior in idealized laboratory tasks (see Dhami, Hertwig, & Hoffrage, 2004). Our results suggest that it does not pay off to tune one's use of exemplar- and rule-based processes to the type of object one wants to generalize to within the same task. This conclusion seems inconsistent with empirical studies suggesting that participants successfully switch between processes in categorization tasks (e.g., Erickson, 2008). Yet these experimental tasks may be unrepresentative of real-world environments. In many experimental studies—especially in categorization research—there is typically little (or no) doubt about which process is better suited to solving the overall task (or responding to a specific object), because there is no irreducible uncertainty (i.e., no noise and thus the relationship between cues and criterion could, in principle, be learned perfectly) and participants are provided with ample learning experience on typically only a small number of cues. Participants can thus easily learn to select between or differentially use the two processes. We speculate that deviating from a simple blending strategy is generally worthwhile only in environments in which one process is clearly superior to the other, both processes make similar errors, and it is possible to ascertain this statistical structure with enough confidence (see Herzog & Hertwig, 2014b; Soll & Larrick, 2009). We would argue, however, that this is typically not the case in real-world environments, because there is usually irreducible uncertainty, and learning experience is often limited. It would thus seem prudent that human judges and decision makers, as modeled, for example, by ATRIUM (Erickson & Kruschke, 1998), start with a simple blend of both processes and deviate from this approach (e.g., by selection or object-specific tuning) only when feedback justifies it. In addition, understanding the complementary strengths and weaknesses of different cognitive processes from a crowd-within perspective (Herzog & Hertwig, 2009; 2014a; 2014b) could offer ways of improving human estimation, categorization, and decision making by boosting decision makers' predictive skills

through instructing them when and how to select or blend exemplar-based and rule-based processes, in particular, and different cognitive processes, in general.

References

Aha, D. (1997). Lazy learning. *Artificial Intelligence Review*, *11*, 7–10. doi:10.1023/A:1006538427943

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*, 629–647. doi:10.3758/BF03196200

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036–1060. doi:10.1037/0033-295X.111.4.1036

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic Publishers.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning, *105*, 442–481. doi:10.1037/0033-295X.105.3.442

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*, 137–154. doi:10.1016/0001-6918(94)90048-5

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*, 1772–1784. doi:10.1016/j.jbusres.2015.01.061

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*, 2–16. doi:10.1016/j.cognition.2010.10.004

Bröder, A., & Schiffer, S. (2003). Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General*, *132*, 277–293. doi:10.1037/0096-3445.132.2.277

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making*, *19*, 361–380. doi:10.1002/bdm.533

Bröder, A., Newell, B. R., & Platzer, C. (2010). Cue integration vs. exemplar-based reasoning in multi-attribute decisions from memory: A matter of cue representation. *Judgment and Decision Making*, *5*, 326–338. Retrieved from http://journal.sjdm.org/10/10614a/jdm10614a.html

Brooks, L. R., & Hannah, S. D. (2006). Instantiated features and the use of "rules." *Journal of Experimental Psychology: General*, *135*, 133–151. doi:10.1037/0096-3445.135.2.133

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, *6*, 5–20. doi:10.1016/j.inffus.2004.04.004

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*, 57–65. doi:10.1016/S1364-6613(98)01273-X

Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, *29*, 317–331. doi:10.3102/10769986029003317

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*, 79–101. doi:10.1037/dec0000004

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988. doi:10.1037/0033-2909.130.6.959

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, *1857*, 1–15. doi:10.1007/3-540-45014-9_1

Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulus-dependent representation. *Memory & Cognition*, *36*, 749–761. doi:10.3758/MC.36.4.749

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140. doi:10.1037/0096-3445.127.2.107

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58. doi:10.1162/neco.1992.4.1.1

Geurts, P. (2010). Bias vs variance decomposition for regression and classification. In O. Maimon & R. Lior (Eds.), *Data mining and knowledge discovery handbook* (pp. 733–746). New York, NY: Springer. doi:10.1007/978-0-387-09823-4_37

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. doi:10.1037/0033-295X.103.4.650

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford, United Kingdom: Oxford University Press.

Gigerenzer, G., Todd, P. M., ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decision*, *15*, 261–278. doi:10.1007/BF00125672

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, *65*, 197–230. doi:10.1016/S0010-0277(97)00044-9

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, *114*, 1–18. doi:10.1016/j.cognition.2009.08.011

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection. *Mathematical Proceedings of the Cambridge Philosophical Society*, *23*, 607–615. doi:10.1017/S0305004100011750

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. New York, NY: Oxford University Press.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*, 494–508. doi:10.1037/0033-295X.112.2.494

Hertwig, R., Hoffrage, U., & Sparr, R. (2012). How estimation can benefit from an imbalanced world. In P. M. Todd & G. Gigerenzer (Eds.), *Ecological rationality: Intelligence in the world* (pp. 379–406). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195315448.003.0116

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*, 504–506. doi:10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 218–232. doi:10.1037/a0034054

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science*, *24*, 869–879. doi:10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*, 2242–2261. doi:10.1037/a0037989

Homa, D., Proulx, M. J., & Blair, M. (2008). The modulating influence of category size on the classification of exception patterns. *The Quarterly Journal of Experimental Psychology*, *61*, 425–443. doi:10.1080/17470210701238883

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688. doi:10.1016/j.ijforecast.2006.03.001

Jacobs, R. A. (1999). Computational studies of the development of functionally specialized neural modules. *Trends in Cognitive Sciences*, *3*, 31–38. doi:10.1016/S1364-6613(98)01260-1

Jacobs, R. A., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87. doi:10.1162/neco.1991.3.1.79

Johansen, M., & Palmeri, T. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553. doi:10.1016/S0010-0285(02)00505-4

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607. doi:10.1207/s15516709cog2605_2

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 259–298. doi:10.1016/j.cognition.2007.02.003

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review*, *14*, 1140–1146. doi:10.3758/BF03193103

Kelley, H., & Busemeyer, J. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology*, *52*, 218–240. doi:10.1016/j.jmp.2008.01.009

Krause, J., Ruxton, G. D., & Krause, S. (2010). Swarm intelligence in animals and humans. *Trends in Ecology and Evolution*, *25*, 28–34. doi:10.1016/j.tree.2009.06.016

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi:10.1037/0033-295X.99.1.22

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 267–301). New York, NY: Cambridge University Press.

Kuncheva, L. (2004). *Combining pattern classifiers: Methods and algorithms*. Hoboken, NJ: John Wiley & Sons.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*, 111–127. doi:10.1287/mnsc.1050.0459

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227–242). New York, NY: Psychology Press.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition*, *39*, 914–923. doi:10.3758/s13421-010-0059-7

Luan, S., Katsikopoulos, K. V., & Reimer, T. (2012). When does diversity trump ability (and vice versa) in group decision making? A simulation study. *PLoS ONE*, *7*, e31043. doi:10.1371/journal.pone.0031043.s001

Luce, D. (1959). *Invidual choice behavior*. New York, NY: Wiley.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*, 276–299. doi:10.1037/a0036677

Marling, C., Sqalli, M., Rissland, E. L., Munoz-Avila, H., & Aha, D. (2002). Case-based reasoning integrations. *AI Magazine*, *23*, 69–86. doi:10.1609/aimag.v23i1.1610

Marewski, J. N., & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*, 39–59. doi:10.1002/wcs.1265

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 676–692. doi:10.1037/0096-1523.34.3.676

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *The Quarterly Journal of Experimental Psychology*, *62*, 890–908. doi:10.1080/17470210802351411

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M., & Johansen, M. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375–402.

Nosofsky, R. M., & Zaki, S. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*, 247–255. doi:10.1111/1467-9280.00051

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79. doi:10.1037/0033-295X.101.1.53

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*, 207–240. doi:10.1016/j.cogpsych.2012.03.003

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, United

Kingdom: Cambridge University Press.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, *4*, 211–255. doi:10.1162/153244304322972694

Platzer, C., & Bröder, A. (2013). When the rule is ruled out: Exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*, 429–441. doi:10.1002/bdm.1776

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*, 1–14. doi:10.1017/S0140525X05000014

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 811–829. doi:10.1037/0278-7393.31.5.811

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236. doi:10.1037/0096-3445.135.2.207

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar- and rule-based theories of categorization. *Current Directions In Psychological Science*, *15*, 9–13. doi:10.1111/j.0963-7214.2006.00397.x

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*, 39–64. doi:10.1037/a0030777

Schwarz, G. (1978). Estimating the dimension of a model. *Annals Of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*, 317–346. doi:10.1006/cogp.1998.0699

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 780–805. doi:10.1037/a0015145

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, *146*, 84–96. doi:10.1016/j.actpsy.2013.12.007

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*, 266–298. doi:10.1037/a0018620

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how*

collective wisdom shapes business, economies, societies and nations. Garden City, NY: Doubleday.

Thomson, R., Lebiere, C., Anderson, J. R., & Staszewski, J. (2015). A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition, 4*, 180–190. doi:10.1016/j.jarmac.2014.06.002

Timmermann, A. G. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. G. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 135–196). Amsterdam, Netherlands: North Holland.

Todd, P. M., Gigerenzer, G., the ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. Oxford, United Kingdom: Oxford University Press.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732–749. doi:10.3758/PBR.15.4.732

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*, 73–96. doi:10.1037/0096-3445.137.1.73

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 867–889. doi:10.1037/a0015501

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a Doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology*, *61*, 12–22. doi:10.1027/1618-3169/a000221

von Helversen, B., Karlsson, L., Mata, R., & Wilke, A. (2013). Why does cue polarity information provide benefits in inference problems? The role of strategy selection and knowledge of cue importance. *Acta Psychologica*, *144*, 73–82. doi:10.1016/j.actpsy.2013.05.007

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647. doi:10.1111/j.1467-9280.2008.02136.x

Table 1. Characteristics of the synthetic data sets

Data set	Criterion (mean)	Criterion (range)	Cue–criterion correlations	R^2	Skew
Linear	61	–18, 147	.47, .38, .34, .25, .13, .12	.45	0.09
Multiplicative	6	–52, 106	.22, .13, .12, .09, .08, .07	.08	0.33
Mixed	33	–52, 147	.25, .19, .16, .15, .06, .05	.12	0.20

Note. N = 1,000; Number of cues = 6; R^2 = percentage of linear variance explained. Skew = Pearson's moment coefficient of skewness of the data set.

Table 2. Cross-validated standardized estimation accuracy of the six coordination schemes in three synthetic data sets

Data set	Strategy	Sample size						
		20	40	60	80	100	200	500
Linear	Rule model	0.917[3]	0.816[2]	**0.788**[1]	**0.778**[1]	**0.769**[1]	**0.757**[1]	**0.748**[1]
	Exemplar model	0.929[3]	0.870[2]	0.845[1]	0.835[1]	0.826[1]	0.805[1]	0.786[1]
	Selection-task	0.945[3]	0.865[2]	0.837[2]	0.823[1]	0.809[1]	0.778[1]	0.752[1]
	Selection-object	0.923[3]	0.825[2]	0.800[1]	0.792[1]	0.783[1]	0.770[1]	0.762[1]
	Blending-task	**0.871**[2]	**0.813**[1]	0.792[1]	0.786[1]	0.779[1]	0.767[1]	0.757[1]
	Blending-object	0.884[3]	**0.811**[1]	0.791[1]	0.784[1]	0.777[1]	0.766[1]	0.757[1]
Mixed	Rule model	1.162[4]	1.034[2]	0.998[1]	0.984[1]	0.977[1]	**0.958**[1]	**0.950**[1]
	Exemplar model	**1.036**[2]	1.002[2]	0.990[1]	0.985[1]	0.981[1]	0.969[1]	0.961[1]
	Selection-task	1.050[3]	1.004[2]	0.990[1]	0.985[1]	0.981[1]	0.969[1]	0.961[1]
	Selection-object	1.161[4]	1.036[2]	1.000[1]	0.986[1]	0.979[1]	0.961[1]	0.954[1]
	Blending-task	1.054[2]	**0.996**[1]	**0.979**[1]	**0.973**[1]	**0.969**[1]	**0.957**[1]	**0.952**[1]
	Blending-object	1.079[3]	1.001[2]	**0.980**[1]	**0.973**[1]	**0.969**[1]	**0.957**[1]	**0.952**[1]
Multiplicative	Rule model	1.193[5]	1.058[2]	1.022[2]	1.005[1]	1.000[1]	0.979[1]	0.970[1]
	Exemplar model	**1.040**[3]	**1.007**[2]	**0.996**[1]	0.989[1]	**0.985**[1]	**0.968**[1]	**0.955**[1]
	Selection-task	1.050[3]	**1.007**[2]	**0.996**[1]	0.989[1]	**0.985**[1]	**0.968**[1]	**0.955**[1]
	Selection-object	1.190[5]	1.054[2]	1.019[2]	1.002[1]	0.996[1]	0.973[1]	0.958[1]
	Blending-task	1.073[3]	1.012[2]	**0.995**[1]	**0.986**[1]	**0.984**[1]	**0.967**[1]	**0.957**[1]
	Blending-object	1.100[3]	1.016[2]	**0.996**[1]	**0.987**[1]	**0.984**[1]	**0.967**[1]	**0.957**[1]

Note. The values show the cross-validated standardized estimation accuracy (RMSSE) of the six coordination schemes in the three synthetic data sets (linear, mixed, and multiplicative) for learning samples of different sizes. The single digits in square brackets indicate the third digit of the standard error of the RMSSE (e.g., "0.971[3]" indicates 0.971±0.003). For each data set and sample size, the strategies performing best (i.e., within 2 standard errors of the best-performing strategy) are highlighted in bold. See main text for details on the coordination schemes and data sets.

Table 3. Characteristics of the real-world data sets (adapted from Table 1 in Dana & Dawes, 2004)

Data set	N	k	**v** Vector	$\varnothing r_{xixj}$	R^2	Skew
Abalone	4,177	7	.63 .58 .56 .56 .54 .50 .42	.89	0.53	1.11
ABC	955	5	.32 .20 .06 .04 .02	.08	0.12	0.33
NFL	3,057	10	.46 .43 .37 .34 .33 .27 .21 .07 .05 .05	.21	0.29	-0.01
NES	1,910	6	.26 .17 .15 .15 .13 .12	.11	0.12	-0.41
WLS	6,385	5	.13 .11 .10 .10 .10	.15	0.04	-0.26

Note. N = number of cases, k = number of cues, **v** Vector = zero-order correlation between target variable and cues, $\varnothing r_{xixj}$ = mean intercorrelation among cues, R^2 = percentage of linear variance explained, Skew = Pearson's moment coefficient of skewness of the data set.

Table 4. Cross-validated standardized estimation accuracy of the six coordination schemes in the five real-world data sets

Data set	Strategy	Sample size						
		20	40	60	80	100	200	500
Summary	Rule model (5/35)	1.155	0.991	0.946	0.928	0.918	0.897	0.886
	Exemplar model (10/35)	**0.995**	0.960	0.943	0.934	0.928	0.912	0.899
	Selection-task (6/35)	1.029	0.962	0.941	0.931	0.924	0.908	0.896
	Selection-object (0/35)	1.151	0.993	0.949	0.932	0.922	0.902	0.890
	Blending-task (29/35)	1.009	**0.941**	**0.920**	**0.911**	**0.905**	**0.892**	**0.884**
	Blending-object (21/35)	1.057	0.949	0.922	0.913	0.906	0.893	**0.884**
Abalone	Rule model	0.999[9]	0.826[5]	0.767[3]	0.750[3]	0.740[3]	0.710[2]	0.699[2]
	Exemplar model	0.888[3]	0.839[3]	0.809[2]	0.794[2]	0.784[2]	0.747[2]	0.724[2]
	Selection-task	0.970[8]	0.846[3]	0.799[3]	0.779[3]	0.764[3]	0.725[2]	0.705[2]
	Selection-object	0.991[9]	0.832[4]	0.777[3]	0.759[3]	0.753[3]	0.721[2]	0.706[2]
	Blending-task	**0.837**[4]	**0.764**[2]	**0.734**[2]	**0.725**[2]	**0.720**[2]	**0.697**[2]	**0.688**[2]
	Blending-object	0.905[6]	0.781[3]	0.741[2]	**0.729**[2]	**0.724**[2]	**0.700**[2]	**0.690**[2]
NFL	Rule model	1.268[8]	0.996[3]	0.936[2]	0.911[2]	0.896[1]	**0.867**[1]	**0.856**[1]
	Exemplar model	**0.972**[2]	0.937[2]	0.920[2]	0.910[1]	0.904[1]	0.890[1]	0.881[1]
	Selection-task	1.019[6]	0.937[2]	0.920[2]	0.910[1]	0.904[1]	0.890[1]	0.880[1]
	Selection-object	1.263[8]	0.996[3]	0.938[2]	0.914[2]	0.900[1]	0.873[1]	0.862[1]
	Blending-task	1.029[3]	**0.925**[2]	**0.900**[1]	**0.888**[1]	**0.881**[1]	**0.866**[1]	0.859[1]
	Blending-object	1.127[6]	0.937[2]	0.903[2]	**0.890**[1]	**0.881**[1]	**0.866**[1]	0.859[1]
ABC	Rule model	1.148[5]	1.021[2]	0.985[1]	0.976[1]	0.967[1]	**0.951**[1]	**0.944**[1]
	Exemplar model	**1.040**[2]	1.006[2]	0.991[1]	0.985[1]	0.979[1]	0.967[1]	0.956[1]
	Selection-task	1.053[3]	1.007[2]	0.992[1]	0.986[1]	0.979[1]	0.967[1]	0.956[1]
	Selection-object	1.146[5]	1.024[2]	0.988[1]	0.979[1]	0.971[1]	0.957[1]	0.949[1]

	Blending-task	1.049[3]	**0.993**[1]	**0.974**[1]	**0.968**[1]	**0.962**[1]	**0.952**[1]	**0.945**[1]
	Blending-object	1.071[3]	0.996[2]	**0.974**[1]	**0.969**[1]	**0.963**[1]	**0.953**[1]	**0.946**[1]
NES	Rule model	1.167[5]	1.041[2]	1.003[2]	0.985[2]	0.973[2]	0.960[1]	**0.946**[1]
	Exemplar model	**1.032**[3]	**1.000**[2]	0.987[2]	0.978[2]	0.971[2]	0.964[1]	0.951[1]
	Selection-task	1.050[4]	**1.001**[2]	0.987[2]	0.978[2]	0.971[2]	0.964[1]	0.951[1]
	Selection-object	1.165[5]	1.042[2]	1.004[2]	0.987[2]	0.974[2]	0.962[1]	0.948[1]
	Blending-task	1.054[3]	**0.999**[2]	**0.980**[2]	**0.970**[2]	**0.962**[1]	**0.956**[1]	**0.945**[1]
	Blending-object	1.082[3]	1.004[2]	**0.982**[2]	**0.971**[2]	**0.963**[1]	**0.956**[1]	**0.945**[1]
WLS	Rule model	1.191[4]	1.069[2]	1.038[2]	1.021[1]	1.012[1]	0.995[1]	0.986[1]
	Exemplar model	**1.044**[2]	**1.019**[1]	**1.008**[1]	**1.004**[1]	**1.000**[1]	0.992[1]	**0.985**[1]
	Selection-task	1.053[3]	**1.020**[2]	**1.008**[1]	**1.004**[1]	**1.000**[1]	0.992[1]	**0.985**[1]
	Selection-object	1.191[4]	1.070[2]	1.038[2]	1.021[1]	1.013[1]	0.995[1]	0.986[1]
	Blending-task	1.078[3]	1.026[1]	**1.010**[1]	**1.003**[1]	**0.999**[1]	**0.989**[1]	**0.983**[1]
	Blending-object	1.101[3]	1.030[2]	1.012[1]	**1.004**[1]	**0.999**[1]	**0.990**[1]	**0.983**[1]

Note. The values show the cross-validated standardized estimation accuracy (RMSSE) of the six coordination schemes in the five real-world data sets (including a summary across data sets) for learning samples of different sizes. The single digits in squared brackets indicate the third digit of the standard error of the RMSSE (e.g., "0.999[9]" indicates 0.999±0.009). For each data set and sample size, the strategies performing best (i.e., within 2 standard errors of the best-performing strategy) are highlighted in bold. In the Summary section, the relative frequencies in parentheses following the strategy names indicate how often each strategy was among the best-performing strategies across the five data sets and seven sizes of learning samples. See main text for details on the coordination schemes and data sets.

Figures

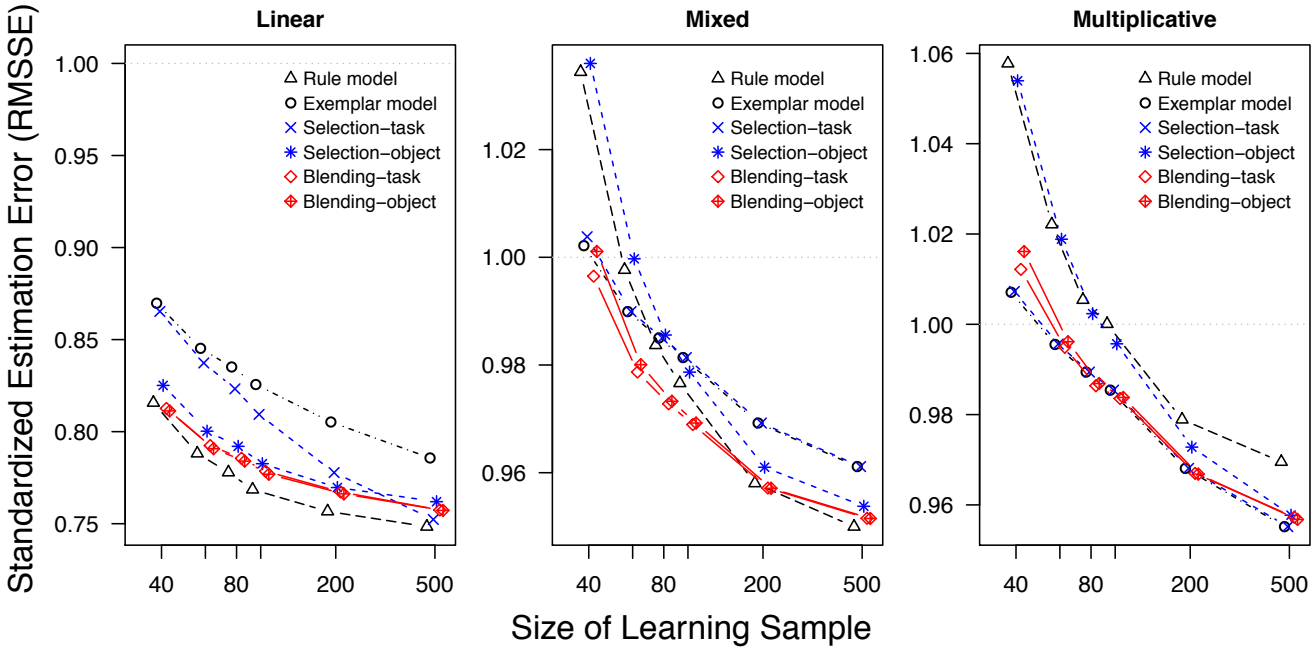

Figure 1. Cross-validated standardized estimation accuracy (RMSSE) of the six coordination schemes in three synthetic data sets (linear, mixed, and multiplicative) for learning samples of increasing size (x-axis scaled by the natural logarithm). The symbols are slightly jittered horizontally to avoid overplotting. Because the scaling of the y-axis differs across panels, we include a horizontal, dotted line at y = 1 to facilitate comparisons across data sets.

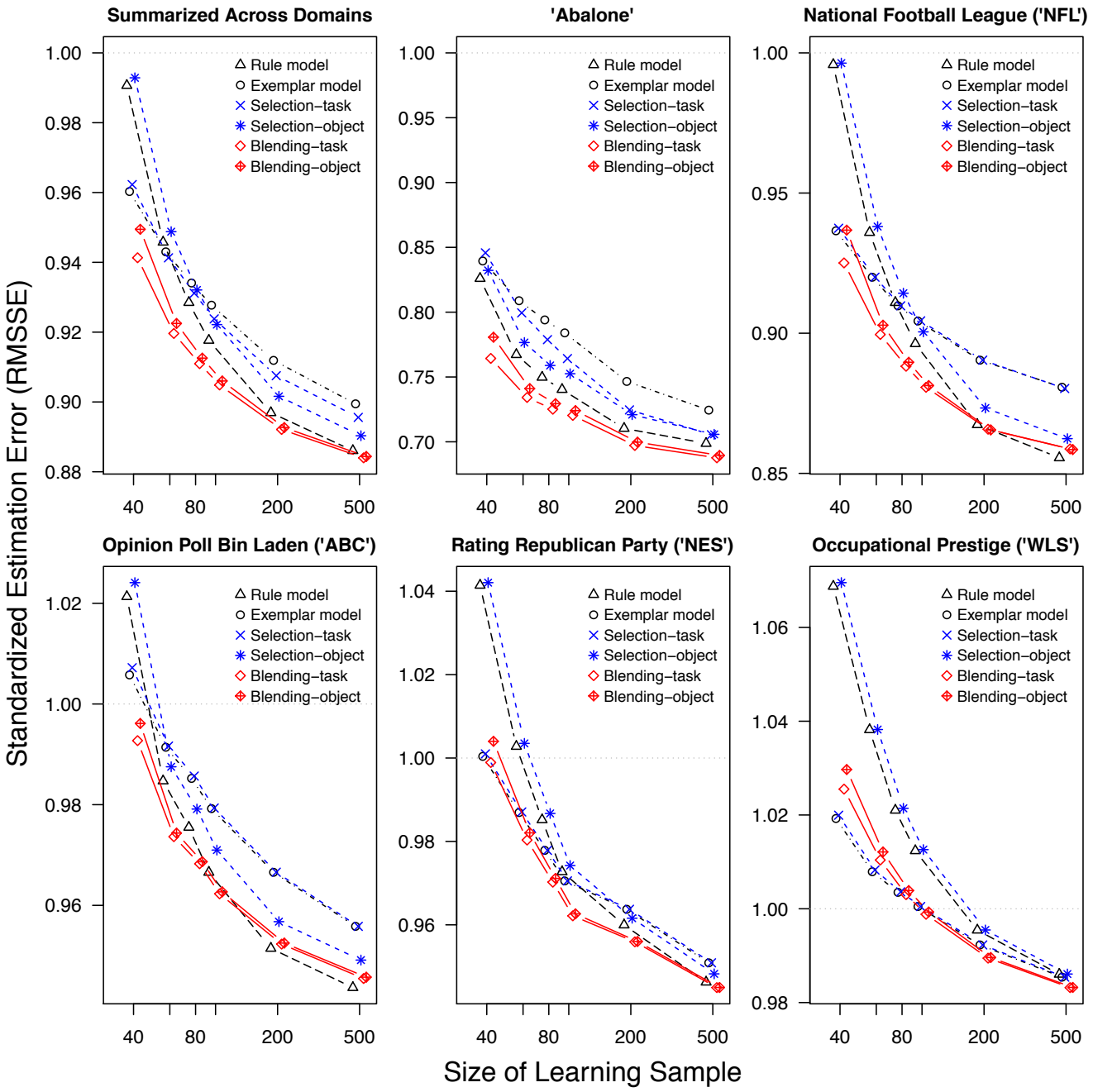

Figure 2. Cross-validated standardized estimation accuracy (RMSSE) of the six coordination schemes in five data sets for learning samples of increasing size (x-axis scaled by the natural logarithm). The summary panel (upper left) averages the RMSSE values across data sets. The points are slightly jittered horizontally to avoid overplotting. Because the scaling of the y-axis differs across panels, we include a horizontal, dotted line at y = 1 to facilitate comparisons across data sets.

Authors' Biographies

Stefan Herzog is a researcher at the Center for Adaptive Rationality (ARC) at the Max Planck Institute for Human Development in Berlin. He is interested in how to improve people's judgment and decision making—with a focus on how to create the "wisdom of crowds" within one mind. His broader research interests are judgment and decision making, bounded rationality and heuristics, social decision making and the wisdom of crowds, medical decision making, cognitive science, decision aids and machine learning.

Bettina von Helversen is Professor for Cognitive Decision Psychology at the Department of Psychology at the University of Zurich. She is interested in how to model people's judgment and decision making—with a focus on strategy selection in decision making, rule-based, similarity-based, and exemplar-based processes, and the role of affect and stress. Her broader research interests are judgment and decision making, bounded rationality and heuristics, development, and cognitive science.